THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 60 September 1981 Number 7, Part 1

Sampling From Structured Populations: Some Issues and Answers

By V. N. NAIR and T. E. DALENIUS*

(Manuscript received February 25, 1981)

This paper reviews some sampling issues that are common to many Bell System surveys. We discuss various aspects of two-stage sampling designs, and emphasize sampling from populations with multiple characteristics. The hierarchical structure of the population in many surveys makes the use of multistage sampling techniques attractive. In populations with multiple characteristics, often not every characteristic is common to every unit. We consider some special designs for sampling from such populations. Finally, we discuss some issues in network sampling. Two recent Bell System surveys are used to illustrate most of the ideas discussed. One of the surveys deals with the estimation of traffic characteristics for various classes of service, while the other one is a survey of baseband transmission impairments.

I. INTRODUCTION

Sample surveys have played an increasingly important role in the Bell System in recent years as a means of providing an objective basis for decision making. To an extent, this has been due to the growing awareness among users of the survey results that, in most surveys, sampling is not the only source of error and often not the primary source. Even if a presumably complete census were taken instead of a sample, serious errors might exist in the results arising from various causes such as measurement or response errors.

* Brown University.

The growth in numbers, in recent years, has also been accompanied by a widening of the range (both in type and complexity) of the surveys. For many of these surveys, a simple and readily available sampling design can easily be adapted to the needs of the prevailing situation. More often, however, the problem at hand is sufficiently complex and nonstandard so that various parts of existing sampling theory have to be modified and pieced together to arrive at a reasonable solution.

Nevertheless, some sampling issues are common to a number of Bell System surveys. Most of these surveys involve sampling from populations that are highly structured, and any cost-efficient sampling design must take this structure into account. In this paper, we review some sampling issues that arose in two surveys currently under implementation. Both surveys possess some common features as well as features unique to themselves. Since these features are common to a large number of other surveys, an exposition of both the theoretical and practical considerations involved may prove beneficial to other survey practitioners. Let us first consider the two examples.

Example 1. Cost of service traffic usage studies (COSTUS)

The various Bell operating telephone companies (OTCs) carry out these surveys periodically to obtain an objective basis for distributing the traffic-sensitive costs for a jurisdiction, typically a state within an OTC, among its various classes of telephone service. Measurements of three traffic characteristics (busy-hour CCS, busy-hour peg count and 14-day peg count) from the sampled telephone lines are used to calculate the relative magnitudes of the traffic characteristics for each class of service. [CCS is a traditional unit for measuring the usage of channels (it stands for hundred call seconds per hour). Peg count is the number of calls actually handled.] These values are then used as inputs to the "embedded direct costs" analysis, which allocates most traffic-sensitive investments and expenses among the various classes of service.

The elementary units in this study are telephone lines corresponding to the various classes of service. These units, however, are clustered into central offices. In fact, each central office has a number of clusters associated with it, one cluster for each class of service. A reasonably cost-efficient design should take this hierarchical clustering into account, since the major portion of the costs in observing a line arises from visiting the central office and setting up the measuring equipment. Thus, a two-stage sampling design with central offices serving as primary sampling units (PSUs) and telephone lines serving as secondary sampling units (SSUs) seems attractive. This is even more so since the central offices provide service in a number of classes of

service so that from each sampled central office, we can further subsample telephone lines from all the available classes of service.

Hence, COSTUS are examples of the use of a two-stage sampling design for a population with multiple characteristics. The different characteristics here correspond to the different classes of service. The parameters of the sampling design in COSTUS are determined so that the busy-hour CCS parameter for each class is estimated with a prescribed accuracy. One additional complication in these studies is the fact that not all central offices provide service in every available class. In some jurisdictions, there are some classes of service (such as coin) that are provided in only a few offices. The sampling literature refers to this as the problem of "partial variate pattern" (PVP). The presence of PVP causes difficulties in selecting an appropriate sample of central offices for the estimation of the parameters of all the classes of service.

Example 2. Survey of baseband transmission impairments

The aim of this survey, currently under development at Bell Laboratories, is to measure baseband transmission impairments for various trunk facility types. From each sampled trunk, estimates of various impairment characteristics, such as signal to C-notched noise ratio (s/n) and second- and third-order harmonic distortion (R2 and R3) are to be obtained. Although the near (transmitting) and far (receiving) end-drop equipment, in addition to the carrier system, determines the trunk type, it is known from past experience that the contribution from the carrier system is the dominant factor. Thus, we do not consider the influence of the end-drop equipment in this study. Six different measurement characteristics are to be measured from each sampled trunk and the parameters of seven different trunk types are to be estimated.

The elementary unit in this survey is the trunk. While the trunks are again clustered into central offices, this clustering is not unique since one trunk is common to a pair (transmitting and receiving) of central offices. In fact, the structure of the population here resembles a graph (network) with the central offices as nodes and trunks as edges (arcs). This survey is an example of network (graph) sampling (see Ref. 1, for example). In this survey, if we sample a particular trunk, we have to visit the pair of end offices connected to the trunk to set up the measuring equipment. This implies that it is cheaper to sample additional trunks connected to those two end offices. Hence, taking the structure of the population into account results in considerable cost savings.

One possible approach to this problem is to use multistage sampling to select pairs of offices and trunks connected to those offices. Since we are interested in different trunk types, this study also involves multiple characteristics.

Both the above examples involve using multistage sampling to study populations with multiple characteristics. Multistage sampling is not an uncommon phenomenon in Bell System surveys where the natural administrative and geographic clustering of units makes it very cost efficient. In Sections II and III we review various issues that confront a survey statistician in developing a two-stage sampling design for studying multiple characteristics. Some of the issues discussed in Section II are also common to other sampling designs. Section III deals primarily with determining the parameters of the sample design. In Section IV, we consider some sampling designs for populations with PVP. Section V is a brief review of issues in network sampling. We conclude the paper with a summary in Section VI. Throughout the paper we try to balance theoretical considerations with practical guidelinea gained from our own experience. One of the two examples is used, wherever possihle, to illustrate the ideas discussed.

II. TWO-STAGE SAMPLING: SOME PRELIMINARIES

This section deals with some preliminary considerations in developing a two-stage sampling design. Some of the discussion deals with issues that are common to sample surveys in general. We hegin with a discussion of the rationale for using two-stage or multistage sampling designs. After an introduction to some notation, we examine how prescrihed accuracy requirements are implemented in a sample survey and discuss the use of prior information. Section 2.6 examines the use of varying prohahility sampling schemes. Section 2.7 discusses ratio estimators with specific emphasis on two-stage sampling situations.

2.1 Why two-stage sampling?

The individuals whose characteristics are to he measured in a study are called elementary units. Observational access to the elementary units, in many cases, is provided hy multistage sampling. Let the elementary units he grouped into a number of suitable clusters. In two-stage sampling, the clusters are used as PSUs and a sample of PSUs is selected in the first stage. The PSUs selected are divided into a numher of SSUs and a sample of SSUs is selected from each PSU selected in the first stage. (The elementary units themselves can serve as SSUs.) All elementary units in the selected SSUs are observed with respect to the variahles of interest.

There are various reasons why multistage sampling is attractive. For instance, in many studies, a complete list ("frame") of elementary units is not available and it may he prohihitively expensive to create such a list. If it is relatively cheap to construct a list of clusters, the clusters can be used as PSUs in a two-stage sampling scheme. Then,

only a list of the elementary units in the sampled clusters needs to be constructed. This results in considerable cost savings.

Often, the population of elementary units in a survey is dispersed over a large geographical area. If we have to visit each sampled unit to collect measurements, sampling from the list of elementary units can lead to high costs per elementary unit. A more cost-efficient scheme may be obtained by grouping the elementary units into geographically compact clusters and using multistage sampling with the clusters as PSUs.

Typically, the cost reduction in multistage sampling is accompanied by an increase in the variance of the estimate over the variance of an estimate from a simple random sampling (SRS) of the same number of elementary units. However, the "accuracy" per unit cost may be higher. If we have some control over the formation of the clusters, we can actually reduce the variance (relative to SRS) by grouping the units so that there is more variation within clusters than between clusters. In most Bell System surveys, however, the clusters are predetermined.

2.2 Notation

We use the following notation throughout the remainder of this paper:

M = number of PSUs in the universe,
m = number of PSUs sampled,
N_i = number of SSUs in PSU i, $i = 1, \cdots, M$,
n_i = number of SSUs selected from the ith sampled PSU, $i = 1, \cdots, m$,
Π_i = probability of selecting the ith PSU in a sample of size m, $\sum_{i=1}^{M} \Pi_i = m$,
Y_{ij} = characteristic to be measured, $j = 1, \cdots, N_i$, $i = 1, \cdots, M$,
y_{ij} = value corresponding to a sample unit, $j = 1, \cdots, n_i$, $i = 1, \cdots, m$,

$$Y_i = \sum_{j=1}^{N_i} Y_{ij} \qquad Y = \sum_{i=1}^{M} Y_i, \qquad \bar{\bar{Y}}_i = Y_i/N_i,$$

$$\bar{\bar{Y}} = Y/N, \qquad N = \sum_{i=1}^{M} N_i, \qquad S_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (Y_{ij} - \bar{\bar{Y}}_i)^2,$$

$$y_i = \sum_{j=1}^{n_i} y_{ij}, \qquad y = \sum_{i=1}^{m} y_i, \qquad \bar{\bar{y}}_i = y_i/n_i,$$

$$\bar{\bar{y}} = y/n, \qquad n = \sum_{i=1}^{m} n_i.$$

We consider only equal probability sampling schemes in stage

two in this paper. The parameter of interest is the overall total $Y = \sum_{i=1}^{M} N_i \bar{\bar{Y}}_i$. $\hat{Y}$ denotes an arbitrary estimator of Y. The same considerations can be used for estimating the average $\bar{\bar{Y}}$ if we rewrite

$$\bar{\bar{Y}} = \sum_{i=1}^{M} W_i \bar{\bar{Y}}_i, \qquad W_i = N_i/N.$$

2.3 Accuracy requirements

The sampling design in a carefully planned survey is determined so that either (*i*) the total cost of the survey is minimized subject to a prescribed requirement on the accuracy of the estimators or (*ii*) the accuracy of the estimators is maximized subject to a constraint on the cost. Since both approaches involve essentially the same considerations (see Section III), let us consider in some detail just the problem of minimizing cost subject to accuracy requirements.

A sampling design, where the units are randomly selected according to given probabilities of selection, permits us to make quantitative statements about the error involved in the estimators. This in turn allows us to determine the sample sizes so that the prescribed accuracy requirements are met. These requirements are typically stated in terms of the error $\mathrm{e} = \hat{Y} - Y$ or some function of the error, $f(\mathrm{e})$, such as relative error, and can be expressed as

$$\Pr\{|f(\mathrm{e})| \leq \delta\} \geq 1 - \alpha \tag{1}$$

for some constants α and δ. In COSTUS, for instance, the sample sizes are determined so that the absolute values of the relative error is less than or equal to 0.1 with probability at least 0.9, i.e., $\alpha = \delta = 0.1$. To implement the accuracy condition (1), large-sample theory is usually used to claim that Y is approximately normally distributed. (It is beyond the scope of this paper to discuss the adequacy of this normal approximation. The interested reader is referred to Refs. 2 to 5.) Equation (1) is equivalent to an expression of an upper bound on the variance [or mean-square error (mse) if $\hat{Y}$ is biased] of $\hat{Y}$.

When estimating several parameters, as in populations with multiple characteristics, we may require that several accuracy criteria be satisfied simultaneously. By using normal approximations, we can state this problem, in general, as minimizing the total cost of the survey subject to a constraint on the variances (or mse's) of the form

$$A\mathbf{v} \leq \gamma, \tag{2}$$

where $\mathbf{v} = (v_1, \cdots, v_p)^T$ is the vector of variances (mse's) of the p estimators, A is a $k \times p$ matrix that specifies the k specific linear combinations of the variances that have to meet the accuracy conditions, and $\gamma = (\gamma_1, \cdots, \gamma_k)^T$ represents the bounds on the accuracies.

For example, if $k = p$ and A is the identity matrix, then all the p parameters need to be estimated with prescribed accuracy. If $k = 1$, then only one particular linear combination of the variances is needed to satisfy an accuracy criterion.

2.4 Variance components

Since the accuracy specifications can be stated in terms of the variances of the individual estimators, we need to examine the components of the variance of the estimator in a two-stage sampling scheme. This will aid us later (Section III) in determining the relative contributions to the variance from stages one and two and the tradeoffs in increasing the sample size in stage one versus that in stage two. If we restrict our attention to linear estimators of the form $\hat{Y}_\alpha = \sum_i \alpha_i \bar{\bar{y}}_i$ for estimating Y, we see that α_i must equal N_i/Π_i for the estimator to be unbiased. With this choice of α, $\hat{Y}_\alpha$ is the well-known Horvitz-Thompson (H-T) estimator.[6] A discussion of some of the properties of this estimator can be found in Ref. 7. Let us restrict our attention to the H-T estimator and examine its variance.

If we select m PSUs with replacement (WR) in stage one with inclusion probabilities Π_i, we have a multinomial sample of size m with success probabilities $Z_i = \Pi_i/m$. If the second-stage units are chosen without replacement, the variance of

$$\hat{Y} = \frac{1}{m} \sum_{i=1}^{m} \frac{N_i}{Z_i} \bar{\bar{y}}_i$$

can be written as the sum of two components:[7-9]

(i) the within-PSU variation W is

$$W = \frac{1}{m} \sum_{i=1}^{M} \frac{N_i^2}{Z_i} \frac{S_i^2}{n_i} (1 - f_{2i}), \tag{3}$$

and

(ii) the between-PSU variation B is

$$B = \frac{1}{m} \sum_{i=1}^{M} Z_i (Y_i/Z_i - Y)^2.$$

Here,

$$S_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (Y_{ij} - \bar{\bar{Y}}_i)^2,$$

the within cluster variance and $1 - f_{2i} = (N_i - n_i)/N_i$, the finite population correction.

If the sampling is done without replacement (WOR) in stage one with varying selection probabilities, the within-PSU variation remains the same. The between-PSU variation, however, depends on second-order

inclusion probabilities which are extremely hard to calculate.[7,10] Hartley and Rao provide some approximations.[10] One possible approximation is, of course, the use of eq. (3), valid for the WR scheme, in the WOR situation. If the sampling fraction m/M is large (say >0.25), this approximation may be unreasonable. When the sampling is done WOR with equal selection probabilities in stage one, i.e., SRSWOR, the B component is given by

$$B = \frac{M^2(1-f)}{m(M-1)} \sum_{i=1}^{M} (Y_i - \bar{Y})^2,$$

where $\bar{Y} = \frac{1}{M}\sum_{i=1}^{M} Y_i$ and $f = m/M$.

For a discussion of variance estimation in two-stage sampling, see Refs. 7, 8, or 9, for example. Some approximate but "quick and easy" methods of variance estimation are discussed in Refs. 11 and 12. If the variance estimator is intended only to provide a rough guide as to the accuracy of the estimator, an approximate, but quick and easy, method is adequate. If the accuracy of the estimator is of great importance and must be demonstrated through the variance estimator, we have to use a "good" variance estimator, such as one with small mse.

2.5 Prior information

We need prior information on the variance of the various estimators and on the sampling costs to determine the sample sizes in a survey. It is rare that we have very good prior information, particularly concerning the variance of the estimators. Preliminary estimates can be obtained from prior surveys or pilot studies. One practice commonly found in the Bell System is the use of data from the entire Bell System to develop preliminary estimates for specific jurisdictions.

To implement the accuracy conditions exactly in a two-stage sampling scheme, we need to know each one of the components of W and B in eq. (3) exactly. Since this is rather unlikely, we usually just use two numbers, one for W and one for B, instead of the individual values for each PSU. These numbers can be interpreted as either the average or the maximum over all PSUs.

When the quality of the prior information is poor (as a consequence of one or more of the above reasons), little can be gained in developing a complex design that may (or may not) be "optimum" for the problem at hand. A simpler design which is less sensitive to the preliminary estimates of the design parameters is more desirable. Also, when the preliminary variance estimators are unreliable, an estimate of the accuracy achieved should always be calculated after the fact from the sample to compare with the prescribed accuracy.

2.6 Varying probability sampling

The sample selection schemes in stages one and two can be based on equal or varying probability sampling techniques. For simplicity, we consider varying probability sampling only in stage one. The considerations here also carry over to other stages. Let us examine how the selection probabilities $\{\Pi_i\}$ should be determined so that the variance of the H–T estimator $\sum_{i=1}^{m} (N_i/\Pi_i)\bar{\bar{y}}_i$, for estimating $Y = \sum_{i=1}^{M} Y_i$, is minimized.

In the simpler situation of one-stage cluster sampling, i.e., $n_i = N_i$, if we take Π_i proportional to Y_i, the variance of the H–T estimator is zero.[7] Hence, if there exists an auxiliary variable X_i which is approximately proportional to Y_i, we can use this auxiliary information to select the Π_i's. In some two-stage sampling situations, we can use the measures of size of the PSU, $\{N_i\}$, to obtain "optimal" selection probabilities. To see this, note that the parameter Y can be written as $\sum_{i=1}^{M} N_i\bar{\bar{Y}}_i$, where $\bar{\bar{Y}}_i$ is the PSU mean, and often the $\bar{\bar{Y}}_i$'s are roughly of the same order of magnitude. In this case, the $Y = N_i\bar{\bar{Y}}_i$ will be roughly proportional to the N_i so that we can take the Π_i proportional to the N_i. This is known as probability proportional to size (PPS) sampling. (In COSTUS, for example, a priori, we expect the average busy-hour CCS per main station to be about the same across central offices.) When sampling from populations with multiple characteristics, there are multiple measures of size, one associated with each characteristic. The optimal selection probabilities are some function of these size measures, depending on the particular accuracy criteria of interest. In addition, there are also cases in which the exact size measures are unknown and we have to use estimated measures.

To develop a cost-efficient design, we need to minimize variance per unit cost rather than the actual variance. The optimal selection probabilities must therefore take the cost structure into account. In COSTUS, where the PSUs are central offices, the sampling costs depend on the type of switching equipment in the office. For example, it is considerably more expensive to visit and set up the measuring equipment in an electronic switching system (ESS) office than in a non-ESS office. If we use formal optimality calculations, we find that with other factors held constant, the optimal selection probability for each PSU is inversely proportional to the square root of the cost of sampling that PSU.[9]

One or more of the above considerations may indicate that even if the PSUs vary greatly in size, the optimal selection probabilities are not too unequal. In such a case, we may be better off using SRS, i.e., equal selection probabilities, since (*i*) the selection scheme is simpler and (*ii*) exact variance formulas are available if, in addition, we are sampling WOR. In some situations, we can actually calculate the gain

from using varying selection probability schemes.[7] If the gain is not substantial in these situations, the use of SRS seems preferable.

Also, even if we use SRS when the PSUs vary greatly in size, we can use ratio estimators, which take into account this variation, to estimate the parameters. This is discussed in Section 2.7.

Finally, we briefly discuss a simple scheme for selecting PSUs with unequal probabilities. Many schemes for unequal probability selection exist,[3,7,10] and, in fact, several procedures may lead to the same inclusion probabilities $\{\Pi_i\}$. The scheme we consider here is for sampling WOR and is known as PPS systematic sampling. Let $\{T_i\}$ denote the cumulative totals of the desired selection probabilities $\{\Pi_i\}$,

$$\sum_{i=1}^{M} \Pi_i = m, \qquad T_i = \sum_{j=1}^{i} \Pi_j.$$

To select m PSUs, first select a random number $u \in [0, 1]$ and then select the m PSUs for which

$$T_{i-1} < u + j \leq T_i, \qquad j = 0, 1, \cdots, m-1.$$

Hartley and Rao consider this procedure with a random arrangement of the PSUs and develop approximate variance expressions for the estimator.[10]

2.7 Use of ratio estimation

So far we have considered only unbiased estimators of the total Y. In some situations we can exploit information available for some auxiliary variable and use a biased estimator, such as the ratio estimator, which has smaller mse than the unbiased estimators. To see this, let $\{X_i\}$ be the known auxiliary variable and let X denote the total corresponding to this variable and $\hat{X}$ denote the estimator of X based on the sample. Since we know the error $\hat{X} - X$, we know how this sample performs in estimating X. Hence, if $\{X_i\}$ and $\{Y_i\}$ are highly correlated, it is intuitively clear that we can improve our original estimator $\hat{Y}$ by exploiting our knowledge of how well the sample estimates X.

The ratio estimator itself is a special case of the general difference estimator $\hat{Y}_a = \hat{Y} + a(\hat{X} - X)$ and is obtained by taking $a = -\hat{Y}/\hat{X}$. This results in the estimator $\tilde{Y} = (\hat{Y}/\hat{X})X$. There are other ways of exploiting the information about $\hat{X} - X$. For instance, a can be a prespecified constant. (If $a = 0$, we get the original estimator $\hat{Y}$ based on the Y measurements alone.) We can also take a to be the regression coefficient $\hat{\beta}$ obtained by regressing the Y_i's on the X_i's.

For the ratio estimator $\tilde{Y}$, the mse of Y can be approximated up to a first-order term by

$$V(\hat{Y}) - 2R\,\mathrm{Cov}(\hat{Y}, \hat{X}) + R^2 V(\hat{X}),$$

where $R = Y/X$ and $\hat{Y}$ and $\hat{X}$ are unbiased estimators of Y and X.[9] Tbus, $\tilde{Y}$ will be more efficient than the unbiased estimator $\hat{Y}$ if $2R\,\mathrm{Cov}(\hat{Y}, \hat{X}) > R^2 V(\hat{X})$. This is likely to be true in practice if the X_i's are appropriately chosen.

In Section 2.6, we saw that in some two-stage sampling situations, the Y_i's are likely to be correlated with the size measures $\{N_i\}$. If PPS sampling is not used (for one or more of the reasons we considered earlier), we can take the N_i's to be the auxiliary variables and use the resulting ratio estimator $\tilde{Y} = \hat{R}N$, where

$$\hat{R} = \sum_{i=1}^{m} \frac{N_i}{\Pi_i} \bar{y}_i \Big/ \sum_{i=1}^{m} \frac{N_i}{\Pi_i}.$$

(If we use PPS sampling, the ratio estimator with the size measure as the auxiliary variable is the same as the unbiased estimator.) Our experience with data from several jurisdictions for COSTUS showed a considerable gain from the use of this ratio estimator.

III. DETERMINING THE DESIGN PARAMETERS

3.1 Cost considerations

The ultimate objective in designing an efficient survey design is the maximization of accuracy per unit cost. To accomplish this, we need to know the cost structure of the survey. We can identify three types of costs in two-stage sampling: (*i*) overhead costs; (*ii*) costs that depend primarily on the number of PSUs in the sample; and (*iii*) costs that depend on the number of SSUs in the sample. Since the overhead costs are fixed, they can be ignored in determining the sample sizes. The costs of sampling PSUs may consist of the costs of selecting, traveling to, locating each sampled PSU, and setting up the measuring equipment. A simple cost function may be of the form

$$mC_1 + m\bar{n}C_2, \tag{4}$$

where C_1 and C_2 are the costs of sampling a PSU and SSU, respectively, and $m\bar{n}$ is the total number of SSUs sampled. Typically, however, the cost functions are more complex. In COSTUS, as we mentioned earlier, the cost of sampling a PSU varies from one PSU to another and depends primarily on the switching equipment in the central office. Further, the cost of sampling a telephone line (SSU) also depends on the switching equipment and so varies from one office to another. There is also a special cost structure in the transmission impairments survey in Example 2. Here, if trunks (edges) are selected by using two-stage sampling to determine the pair of end offices connected to the trunk, it is cost-efficient to select offices with many trunks rather than those with fewer trunks.

3.2 Determining the parameters in a simple situation

Let us consider a simple situation to illustrate the concepts involved in determining the parameters of the optimum design. Suppose the number of SSUs in each PSU is the same and equals $\bar{N}$, the cost function is given by eq. (4) and we use SRSWOR to select the units in botb stages. We need to determine only m, the number of PSUs to be sampled, and $\bar{n}$, the number of SSUs to be sampled from each selected PSU. The variance of tbe H-T estimator can now be written (see Section 2.4) as

$$V(\hat{Y}) = (1 - f_1)\frac{V_1^2}{m} + (1 - f_2)\frac{V_2^2}{m\bar{n}} \tag{5}$$

for some V_1 and V_2. Here $f_1 = m/M$ and $f_2 = \bar{n}/\bar{N}$. A comparison of eq. (5) with the cost function $C = mC_1 + m\bar{n}C_2$ reveals that increases in m and $\bar{n}$ have opposite effects on the variance and costs. Also, it is clear that an increase in m results in greater reduction in the variance than a corresponding increase in $\bar{n}$. Since C_1 is typically much larger than C_2, it is more costly to increase the size of the first stage sample than the size of the second stage sample. All of these factors must be taken into consideration in determining the optimum combination of m and $\bar{n}$.

As mentioned earlier, optimum levels of m and $\bar{n}$ can be determined by minimizing either (*i*) the variance subject to a cost constraint or (*ii*) the cost subject to some accuracy requirements. Both approaches yield essentially the same results. The problem can be formulated mathematically as minimizing a given function subject to a constraint. Standard numerical or analytical techniques (LaGrangian multipliers, Cauchy's inequality) can be used to determine the optimum values of m and $\bar{n}$. In this particular simple situation, explicit expressions for m and $\bar{n}$ can be easily obtained. Suppose we want to minimize the cost subject to the condition that the variance eq. (5) does not exceed some value b. If we can ignore the finite population corrections in eq. (5), the optimum values of $\bar{n}$ and m can be obtained as

$$\bar{n}_{\text{opt}} = \frac{V_2/V_1}{\sqrt{C_2/C_1}}$$

and

$$m_{\text{opt}} = \frac{V_1/\sqrt{C_1}}{A},$$

where

$$A = b/(C_1/V_1^2 + C_2/V_2^2)^{1/2}.$$

The total cost of the survey with these values of m and n is given by

$$C_{\text{opt}} = b/A^2.$$

In practice, one should not be satisfied with just determining the optimum values of m and $\bar{n}$ without examining the behavior of the variance and cost functions near the optimum. Since preliminary estimates of costs and variances may only be approximate, the behavior of these functions in a neighborhood around the optimum should be examined. Relatively flat variance and cost functions near the optimum value indicate robustness against possible moderate errors in the input parameters.

3.3 More general situations and the COSTUS example

In most surveys, the situation is more complex than the one we have just discussed. For example, the PSUs will not necessarily be the same size and the cost function may be more complicated. Even in a general situation, the problem can be formulated in such a way that we can determine, either analytically or numerically, the optimum values of: m, the number of PSUs to be selected; $\{\Pi_i\}$, the inclusion probabilities; and $\{n_i\}$, the number of SSUs to be sampled from each selected PSU. Some of these results for some special cost functions can be found in the literature.[7-9]

We want to emphasize here the importance of simplifying the problem, whenever possible, by using reasonable approximations. In a complex situation where there are too many design parameters to be determined, it is difficult to appreciate the impact of unreliable input values. Reducing the number of parameters through the use of some practical guidelines usually provides us with a better understanding of the problem. We illustrate some of these ideas through the COSTUS example.

The PSUs in COSTUS are central offices and, as mentioned earlier, the cost of sampling the office and telephone lines (SSUs) in the office depends on the type of switching equipment in the office. Since each office provides service in several classes, we have to sample lines from all the available classes in the selected offices. However, not every office provides service in every available class. Since we want to study the parameters of all the classes, we take the first-stage costs of sampling an office with service in only one class to be twice that of an office with service in two classes. Hence, the total costs of the survey can be written as

$$TC = \sum_{i=1}^{m} \left\{ \tilde{D}_{1i} + \sum_{C=1}^{S} D_{2i} n_{Ci} \right\},$$

where $\tilde{D}_{1i} = D_{1i}/\Sigma_i$, D_{1i} = the costs of sampling the ith office, Σ_i = number of classes in the ith office, D_{2i} = costs of sampling a line from

the ith office, and n_{Ci} = number of lines to be selected from the ith office for class C (equals zero if office i does not have service in class C). Since the total cost of this survey is a random quantity, we minimize the expected total cost

$$m\left[\sum_{i=1}^{M} Z_i\left(\tilde{D}_{1i} + D_{2i}\sum_{C=1}^{S} n_{Ci}\right)\right], \tag{6}$$

where $mZ_i = \Pi_i$, the inclusion probabilities.

We need to minimize eq. (6) subject to some accuracy constraints. In this study, the quantity to be estimated is the mean load (in CCS) during the busy hour, $\bar{\bar{Y}}_C$. We require a relative error no larger than 0.1 with probability 0.90 for each of the S classes, $C = 1, \cdots, S$. From Section 2.3, we note that this accuracy can be stated in terms of an upper bound on the mse of the estimator. We use the following approximate expression for the relative mse (rmse) to determine the design parameters:

$$\text{rmse}(\hat{\bar{\bar{Y}}}_C) = \bar{\bar{Y}}_C^{-2}\left[\sum_{i=1}^{M}\frac{W_{Ci}^2}{mZ_i}\left(\frac{S_{Ci}^2}{n_{Ci}} + (\bar{\bar{Y}}_{Ci} - \bar{\bar{Y}}_C)^2\right)\right]. \tag{7}$$

The notation here is the same as in Section 2.2. The additional subscript indicates the class of service. This expression (which in fact equals the relative variance of the unbiased estimator) is the zeroth-order term in the Taylor series expansion for the rmse of the ratio estimator. By not taking into account the higher-order terms which include the correlation between the numerator and denominator of the ratio estimator, this expression, in general, overestimates the variability. However, it is simpler to use and the overestimation may be desirable in view of the unrealiability in preliminary estimates.

Before determining the design parameters, we make two additional simplifications: (*i*) replace $S_{Ci}^2/\bar{\bar{Y}}_C^2$ in the first component of eq. (7) by V_{C2}, a quantity that does not depend on the office i; and (*ii*) replace $(\bar{\bar{Y}}_{Ci} - \bar{\bar{Y}}_C)^2/\bar{\bar{Y}}_C^2$ in the second component by V_{C1}, also a quantity independent of the office. This is reasonable since a priori we do not expect much variation between these values and, in any event, we do not know each one of the individual values. (See also the discussion on prior information in Section 2.5.)

Thus, we want to determine the design parameters which minimize eq. (6) subject to

$$\sum_{i=1}^{M}\frac{W_{Ci}^2}{mZ_i}\left(V_{C1} + \frac{V_{C2}}{n_{Ci}}\right) \le b_C$$

for some b_C, $C = 1, \cdots, S$. Instead of determining m, $\{Z_i\}$ and $\{n_i\}$ from the optimality calculations, we only determine m and $\bar{n}_C$, the average number of SSUs to be sampled from a selected PSU for each

class of service. Once m and $\bar{n}_C$ are determined, we can allocate $m\bar{n}_C$, the total number of lines for class C, to each sampled office inversely in proportion to $(D_{2i})^{1/2}$. We also select $\{Z_i\}$ in advance hy taking them proportional to

$$\{N_{\cdot i}/\mathbf{D}_{1i}{}^{1/2}\},$$

where

$$N_{\cdot i} = \sum_{C=1}^{S} N_{Ci}.$$

Once we substitute these values for $\{n_{Ci}\}$ and $\{Z_i\}$ in the variance and cost functions, it is a relatively easy prohlem to find the values of m and $\bar{n}_C$ that minimize the total expected cost suhject to the accuracy constraints. Since there are only $S + 1$ design parameters involved, it is also easy to examine the behavior of the cost and variance functions near the optimum and investigate the sensitivity to errors in input values.

When COSTUS was implemented in a few jurisdictions, we also examined the advantage gained hy using unequal prohability selection schemes. Since we were using the conservative WR variance formulas for the unequal prohability selection scheme, we found that the loss in "efficiency" from using SRSWOR of offices (with exact variance formulas) was not substantial. This also simplified the computations considerahly.

IV. SAMPLING DESIGNS FOR POPULATIONS WITH PARTIAL VARIATE PATTERNS

4.1 The problem of partial variate pattern (PVP)

A multivariate population (for example, one with multiple characteristics) is said to exhibit a PVP if not all the variates can he ohserved from every unit in the population. In COSTUS, as we noted, not all the central offices provide service in every available class. In the survey of hasehand transmission impairments in Example 2, not all carrier systems appear between each pair of central offices. It is easy to visualize many other studies, both within and outside the Bell System, where the populations exhibit PVP. The prohlem of PVP can he serious if there is great variation in the size of the universe corresponding to each variate. The usual sampling designs may not provide reasonahle assurance that we can select a sample that will allow us to estimate the parameters corresponding to each variate with prescrihed accuracy.

Let us consider some schemes for sampling in the presence of PVP (also see Ref. 13). Since the prohlem of PVP is present in one stage of the selection process only, we restrict our attention to sample selection in the first stage. Thus, suppose there are M units in the population,

of which M_C units have characteristic C, $C = 1, \cdots, S$. Let the sample size, determined by accuracy requirements, for characteristic C be m_C. These sample sizes of course also depend on the particular sampling scheme used.

4.2 Some sampling designs

4.2.1 Modified simple multivariate sampling

Let $m = \max_C m_C$ and suppose we select a sample of $m < M$ units, possibly using different selection probabilities for different units. This is the simple multivariate sampling scheme, intended for populations with no PVP. If $\tilde{m}_C$ denotes the number of sampled units with characteristic C, $\tilde{m}_C$ may be much smaller than m_C and in some cases may even be zero. We can modify this scheme in a number of ways. Instead of selecting $m = \max_C m_C$ units, we can select m^* units, according to selection probabilities $\{\Pi_i\}$, where m^* is determined so that the expected number of units in the sample is at least m_C, $C = 1, \cdots, S$. This can be achieved by taking $m^* = \max_C m_C/p_C$, where p_C is the total of the probabilities $Z_i = \Pi_i/m$ for units with characteristic C. This can be justified if we view the selection of a unit with chracteristic C approximately as a binomial experiment with probability of success p_C. This formulation can alternatively be used to determine m^* such that, say 90 percent of the time, $\tilde{m}_C \geq m_C$, $C = 1, \cdots, S$.

4.2.2 Combined multivariate sampling

Here, we consider S universes, each universe corresponding to the units with characteristic C, $C = 1, \cdots, S$. We select an independent sample of size m_C from each one of the S universes. We then observe every available characteristic from the units selected in all of the S samples. The total number of units selected in these S samples can vary between $\max_C m_C$ and $\sum_{C=1}^{S} m_C$. The main disadvantage of this scheme is that this number may be too large. However, we can exercise some control over this number. One possibility is to give higher selection probabilities to units with more characteristics than those with fewer characteristics (see Section 3.3). Alternatively, instead of selecting m_C units from the universe corresponding to characteristic C, we can select a smaller number, m_C^*, of units. This is because we expect to select some units, in addition to these m_C^* units, with characteristic C from the remaining $S - 1$ samples. So, the number m_C^* can be determined such that either on the average or with prescribed probability, the total number of units with cbaracteristic C exceeds m_C, $C = 1, \cdots, S$. The binomial approximations discussed earlier can be used to determine the $\{m_C^*\}$.

4.2.3 Stratified sampling

We can also try to deal with PVP by stratifying the units so that, within each stratum, the units are internally homogeneous in some sense in terms of the PVP. We consider two stratification techniques here.

In the first scheme, called variate stratification, the strata are determined in terms of the variates (characteristics). Suppose the variates are ordered so that the number of units with variate one is smallest, the number with variate two is next smallest, etc. Then, stratum one consists of all the units with variate one, stratum two consists of all units with variate two and not in stratum one, etc. If we now allocate the total sample size among the strata, we can estimate the parameters corresponding to all the variates, especially the "small" ones. However, this scheme is not foolproof in the sense that it is possible to construct examples where the selected sample does not contain any units with one of the variates.

The second method, pattern stratification, is based on the variate pattern. Here, units with identical variate pattern, i.e., having the same set of characteristics, are grouped into a stratum. Unlike the variate stratification scheme, we can guarantee the required sample size for each variate in this scheme. However, this scheme suffers from the serious drawback that the total sample size may be too large, since the number of different strata (which is smaller than the sample size) can be as large as $\min(M, 2^S - 1)$.

In both these schemes, standard nonlinear programming techniques can be used to determine the sample size for each stratum to minimize cost subject to the variance constraints.

4.2.4 Other methods

It is possible to use sequential sampling schemes to ensure that we select a sample with a given number of units for each characteristic.[13] However, it is extremely difficult to determine analytically the selection probabilities for most of these schemes. One simple sequential method that can be implemented is a two-stage simple multivariate sampling scheme in which a simple multivariate sample is supplemented by a second-stage sample from the remaining units. Although the variance calculations become more involved, they are still tractable.

It also is plausible that ideas from the controlled selection methodology can be applied to the selection of samples from populations with PVP.[14,15,16] However, it is not clear how to characterize explicitly the set of all feasible samples here. Variance calculations also remain a difficult problem with controlled selection.

4.3 The design used in COSTUS

The sampling design used in COSTUS for handling PVP will be described here. As the problem of PVP exists only in the first stage, we consider the selection of units in stage one only.

While examining data from several jurisdictions for the different PVPs, we found that, in most cases, a class of service can be classified as either small or large in terms of the proportion of offices with service in that class. There were very few jurisdictions with medium-sized classes of service.

Since the main concern in the presence of PVP is the ability to estimate parameters corresponding to the small classes of service, we decided to group all offices with services in these classes in stratum 1. A combined multivariate sampling scheme, which guarantees the required sample size from each class, is used to select offices from this stratum. Since the total number of offices sampled under this scheme may be large, we restrict tbe size of this stratum to be no larger than 25 percent of the universe.

We can use a simple multivariate sampling scheme to select a sample from the remaining offices. However, we first identify those classes with service in less than 50 percent of the remaining offices. The offices with service in these classes (and not in stratum 1) are grouped into stratum 2. The remaining offices are grouped into stratum 3. Simple multivariate sampling schemes are then used to select units in strata 2 and 3. By doing this, we have reasonable assurance that the sample sizes for the classes that characterize stratum 2 are not too small compared to the required sizes.

Hence, we see that the sampling design for COSTUS is in fact a three-stage sampling design. In the first stage, the offices are grouped into three strata. Different sampling schemes are used in the different strata to select offices in the second stage. From each office selected in the second stage, telephone lines corresponding to each available class are selected in the third stage.

The design we have used here for handling PVP incorporates specific features of some of the schemes discussed in Section 4.2. The stratification is based on considerations similar to those in the variate stratification scheme. It is, however, adaptive in the sense that it depends on the variate pattern in eacb universe. In our applications, we found that in many jurisdictions stratum 2 was empty and in some situations, where the problem of PVP is not serious, stratum 1 was empty.

We arrived at the final design used in COSTUS by examining data from various jurisdictions for the different types of PVP to expect. This design, while not foolproof, provides a reasonable, practical solution to the problem at hand.

V. SAMPLING FROM NETWORKS

In most surveys, we can treat the population under study merely as a collection of elementary units with no importance attributed to the interrelationships that exist among the units. In some situations, however, these relationships cannot be ignored and the selection of the sample is necessarily affected by the network of relationships that exist in the population. In this section, we briefly review some aspects of network sampling and discuss the sampling design used in Example 2.

5.1 Networks

There are a wide range of surveys in the Bell System tbat deal with sampling from a network. Besides communication networks, network sampling also occurs in studies of other types of traffic flow and transportation facilities. A contact network or sociogram may represent the interrelationships among a group of individuals, households, customers, etc. Other examples include similarity or dissimilarity structures in cluster analysis and multidimensional scaling, where we want to compare a set of objects and group them into classes of similar objects.

A network can be described in abstract terms with the aid of graph theory. An undirected graph (network) consists of a nonempty set V of elements called vertices (nodes) and a set of E of elements called edges. Each edge e of E is associated with a pair of vertices (i, j). The edges may have several attributes associated with them. A network can also be represented by a matrix with the columns and rows representing the vertices. A one in the (i, j)th cell of the matrix indicates that the vertices i and j are connected. In the survey of baseband transmission impairments discussed in Example 2, the vertices are central offices and the edges are trunks. In this case, there are many trunks and also different types of trunks between a pair of central offices. Several attributes, corresponding to the impairment characteristics, are associated with each trunk.

5.2 Some sampling schemes

The manner in which we have observational access to the elementary units is the key to developing a reasonable sampling design. If we have a "frame" of all tbe edges in the graph from which we can select a sample of units, the problem is essentially one in traditional sampling theory. If no such frame is available and the structure of the relationship between the nodes must be discovered and explored during the course of data collection, the sampling design problem is quite different. Even in cases in which a complete listing of the edges is available, as in Example 2, cost considerations may dictate that a sample of

edges be selected by first sampling the nodes. Also, unlike traditional sampling where information about a unit can be obtained only by sampling and observing it, information about the relationship between several nodes may be obtained at any one of the nodes in network sampling.

The field of sampling from networks bas been considered by only a few authors so far.[1,17-22] Most of the attention bas been focused on surveys for which the structure of interrelationships is unknown and must be discovered. The references above deal mainly witb estimating parameters that measure various aspects of these relationships.

Goodman proposed the "snowball" sampling scheme for selecting edges (or pairs of connected nodes).[20] In this procedure, the survey proceeds from an initial sample of nodes by obtaining information about other nodes to which they are connected. The next step is to add to the sample some or all of these connected nodes, obtaining data from them as well as information about still other nodes to which they are connected. In an s-stage k-name snowball sample, this process is repeated for s stages and at each stage, k other nodes connected to a node already in the sample are selected. Goodman studies this scheme in detail under the assumption that the initial sample is selected through binomial sampling.[20] He also considers the case in which the k nodes are selected randomly at each stage. See also Ref. 1.

To consider two other methods of network sampling, let us view the network as a matrix with the vertices corresponding to the columns and rows and the elements of the matrix corresponding to the edges. If we select a sample of nodes (rows/columns of the matrix), we can base our inference entirely on the sampled subnetwork that corresponds to the sampled rows and columns. This procedure (called subnetwork sampling) of selecting one or even several subsystems out of a number of subsystems is equivalent to traditional one-stage cluster sampling. It leaves open all questions about interrelationships between one cluster and another. In the partial network sampling scheme, we select a sample of nodes from the node set, and observe all the edges connected to one or more of the nodes in the sample. Estimation of the network characteristics using these two schemes is discussed in Refs. 1 and 18.

5.3 Survey of baseband transmission impairments

In this survey, there are a number of trunks of various types with each trunk associated with a pair of end offices. If we select a particular pair of end offices, it then becomes cheaper to select additional trunks from those trunks that terminate in either one of the two offices. This special cost structure implies that we need to select trunks (edges) by appropriately selecting offices (nodes) to which they are connected.

A multistage sampling scheme is used in this survey. A sample of primary offices, using probabilities proportional to some measure of size (the number of trunks), is selected in the first stage. A number of secondary offices are selected, again using probabilities proportional to some measure of size, from the set of offices connected to each of the primary offices. From every pair of end offices thus sampled, a number of trunks corresponding to each trunk type are selected using simple random sampling. The parameters of the sampling design (m, the number of primary offices, $\{m_i\}$, tbe number of secondary offices and $\{n_{ij}\}$, the number of trunks of a particular type) can all be determined so that the total survey cost is minimized subject to some accuracy criterion.

The two-stage sampling scheme used here to select the pair of end offices can also be viewed as a two-stage snowball sampling scheme. It is of course possible to use a k-stage snowball sample to select the offices. Optimality considerations relating to the number of stages and the sample size in a snowball sample have yet to be resolved.

VI. SUMMARY

We have reviewed various aspects of sampling from structured populations in this paper. The issues that have been selected for discussion, two-stage sampling from populations with multiple characteristics and sampling designs for populations with PVP and network sampling, are common to many Bell System surveys. Thus, we hope that an exposition of some of the theoretical and practical considerations involved in dealing with these situations will serve other survey practitioners. Throughout the paper we have tried to balance theoretical considerations with practical guidelines gained from our own experience. Two recent Bell System surveys are used to illustrate the ideas disscused.

VII. ACKNOWLEDGMENTS

We are grateful to R. E. Hausman and C. H. Morton for their kind permission to use the baseband transmission impairments survey as an example. We have also benefited from the earlier experiences of M. T. Chao and F. W. Nolte in the Cost of Service Traffic Usage Studies.

REFERENCES

1. O. Frank, "Survey Sampling in Graphs," J. Statist. Planning Inference, *1* (1977), pp. 235–64.
2. J. Hajék, "Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population," Ann. Math. Statist., *35* (1964), pp. 1491–523.
3. J. N. K. Rao, "Sampling Designs Involving Unequal Probabilities of Selection and Robust Estimation of a Finite Population Total," in *Contributions to Survey Sampling and Applied Statistics*, edited by H. A. David, New York: Academic, 1978.

4. B. Rosén, "Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I and II," Ann. Math. Statist., *43* (1972), pp. 373–97, 748–76.
5. B. Rosén, "Asymptotic Theory for Des Raj's Estimator, I and II," Scand. J. Statist., *1* (174), pp. 71–83, 135–44.
6. D. G . Horvitz and D. J. Thompson, "A Generalization of Sampling Without Replacement from a Finite Universe," J. Am. Statist. Assoc., *47* (1952), pp. 663–85.
7. M. N. Murthy, *Sampling Theory and Methods*, Calcutta: Statistical Publ. Soc., 1977.
8. W. G. Cochran, *Sampling Techniques*, New York: Wiley, 1977, 3rd ed.
9. M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*, Vols. I and II, New York: Wiley, 1953.
10. H. O. Hartley and J. N. K. Rao, "Sampling with Unequal Probabilities Without Replacement," Ann. Math. Statist., *33* (1962), pp. 350–74.
11. D. R. Brillinger, "Approximate Estimation of the Standard Errors of Complex Statistics Based on Sample Surveys," New Zealand Statist., *II*, No. 2 (1976), pp. 35–41.
12. B. V. Shah, "Variance Estimates for Complex Statistics from Multistage Sample Surveys," in *Survey Sampling and Measurement*, edited by N. K. Namboodiri, New York: Academic, 1978.
13. T. E. Dalenius and O. Frank, "Sampling Populations with Partial Variate Patterns," Scand. J. Statist., *1* (1974), pp. 19–27.
14. R. S. Cochran, "Sampling in Two or More Dimensions," in *Contributions to Survey Sampling and Applied Statistics*, edited by H. A. David, New York: Academic, 1978.
15. L. A. Goodman and L. Kish, "Controlled Selection—A Technique in Probability Sampling," J. Am. Statist. Assoc., *45* (1950), pp. 330–72.
16. R. J. Jessen, "Probability Sampling with Marginal Constraints," J. Am. Statist. Assoc., *65* (1970), 776–96.
17. A. R. Bloemena, *Sampling from a Graph*, Amsterdam: Mathematisch Centrum, 1964.
18. O. Frank, *Statistical Inference in Graphs*, Stockholm: Försvarets Forskningsanstalt, 1971.
19. O. Frank, "A Note on Bernoulli Sampling in Graphs and Horvitz–Thompson Estimation," Scand. J. Statist., *4* (1977), pp. 178–80.
20. L. A. Goodman, "Snowball Sampling," Ann. Math. Statist., *32* (1961), pp. 148–70.
21. M. Granovetter, "Network Sampling: Some First Steps," Am. J. Sociol., *81* (1976), pp. 1287–302.
22. F. F. Stephan, "Three Extensions of Sample Survey Technique: Hybrid, Nexus, and Graduated Sampling," in *New Developments in Survey Sampling*, edited by N. L. Johnson and H. Smith, New York: Wiley, 1969.